结合基因遗传和贪婪搜索的布谷鸟社区检测算法 *

王小刚，闫光辉，周　宁

(兰州交通大学 电子与信息工程学院, 兰州 730070)

摘　要：为了提高复杂网络社区结构挖掘的精度，结合基因遗传和贪婪搜索提出一种面向模块度优化的布谷鸟社区检测算法(GGCSCA)。布谷鸟种群在有序邻居表上逐维随机游走，并采用优质基因遗传策略，使得种群高效优化，同时应用局部模块度增量最大化的贪婪偏好搜索算法快速提升种群质量，以取得好的社区划分结果。GGCSCA 在基准网络和经典网络上进行了实验，并与一些典型算法进行对比，结果说明了本社区发现算法的有效性、准确性和快速收敛性，具有较强的社区识别能力，能够精细地检测出网络社区结构。

关键词：复杂网络；网络社区；布谷鸟搜索算法；贪婪搜索；基因遗传

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.08.0871

Cuckoo search algorithm combining gene inheritance and greedy search for community detection

Wang Xiaogang, Yan Guanghui, Zhou Ning

(*School of Electronic & Information Engineering*, *Lanzhou Jiaotong University*, *Lanzhou* 730070, *China*)

Abstract: In order to improve the accuracy of community detection for complex networks, this paper proposed an algorithm based on cuckoo search algorithm combining gene inheritance and greedy search (GGCSCA) to optimize modularity for community detection. Cuckoos walked randomly on ordered adjacent table and employed gene inheritance strategy, which aim to optimize population efficiently. The algorithm improved population quality quickly by greedy preference search of local modularity increment maximum for the purpose of getting good result of community partition. GGCSCA has been tested on both benchmark networks and some typical complex networks, and compared with some typical community detection algorithms. Experimental results show the effectiveness, accuracy and fast convergence of this algorithm for discovering community structure.　It has strong capability of community identification and can detect the structure of community finely.

Key Words: complex network; network community; cuckoo search algorithm; greedy search; gene inheritance

0　引言

复杂网络是对复杂系统重要特征的抽象概括，其核心思想是将现实系统中所有实体及实体间的关系转换为网络的节点和边，以网络的形式来描述系统中各部分之间的关系，便于深入分析系统结构的拓扑特性，揭示现实系统的本质规律。很多现实系统可以表示为复杂网络或转换为复杂网络，如社交网络、生物网络、计算机网络、交通网络等，复杂网络已经是重要的多学科交叉研究领域。

复杂网络由若干社区组成，社区内部的连接比较紧密，而社区之间的连接比较松散[1]。社区发现探测复杂网络中固有的社区结构，是当前复杂网络领域的一个研究热点。社区发现有助于分析复杂网络功能、内在规律和拓扑结构特性，提供复杂网络演化研究的中观视角，其研究成果已被成功应用于蛋白质功能预测、恐怖组织识别、舆情分析、客户关系聚类、搜索引擎等多方面[2-3]。

复杂网络社区挖掘已经获得了广泛的研究，涌现了很多社区检测算法。常见典型算法主要有基于图分解、分裂的算法，如 Kernighan-Lin 算法[2]，谱分法[4]，GN 算法[1]；基于凝聚的算法，如 Fast Newman 算法[5]，CNM 算法[6]；基于随机游走的方法，如 Walk Trap[7]算法；基于优化的算法，这种方法通过对某种社区评价函数进行优化得到划分结果，常见的是基于模块度指标的优化算法。Newman 和 Girvan 定义了模块度函数[8]，用于评价社区划分。基于模块度优化的算法如模拟退火算法[9]、BGLL 算法[10]、极值优化[11]等。虽然有了很多的社区探测算法，但是社区识别的准确性、效率、易用性甚至通用性等方面还需改善。

基金项目：国家自然科学基金资助项目（61163010，61650207）；甘肃省科技计划资助项目（1610RJZA059）；兰州市科技计划项目（2014-1-171）

作者简介：王小刚（1976-），男，甘肃榆中人，副教授，博士研究生，主要研究方向为软件工程、复杂网络（reswxg@mail.lzjtu.cn）；闫光辉（1970-），男，教授，博士，主要研究方向为数据挖掘，复杂网络；周宁（1979-），男，副教授，博士，主要研究方向为形式化验证.

1 相关工作

近些年，通过智能进化算法进行社区检测已经成为热点。选择合适的社区划分评价函数，可以将复杂网络社区发现问题转换为优化问题，但是实现最优化目标往往是 NP 难的问题。利用智能进化算法，选择合适的启发式规则可以取得较好的近似优化结果。

智能进化社区发现算法主要分为多目标优化和单目标优化。多目标优化方法如：MOCD-PSO[12]，MDCL[13]。单目标方法一般针对模块度进行优化，基本的思想是通过迭代进化追求模块度的最大化。智能进化算法用于社团检测出现较早的是模拟退火算法，Guimera 等人提出了以模块度为优化函数基于模拟退火算法的复杂网络社区检测算法 GA[9]，该成果于 2005 年被《Nature》报道。

黄发良提出一种基于粒子群优化的网络社区发现算法(CDPSO)[14]，该方法基于节点邻居有序表的编码进行全局搜索，一定程度上缓解了基于二值编码的迭代二划分策略导致的局部最优划分问题，这种基于节点邻居有序的方式在很多进化社区检测算法中得以使用。邱晓辉面向社区发现的改进粒子群优化算法[15]，在上述方法的基础上，使用最多邻居从属的变异策略，即节点以一定概率变异为邻居最多从属的社区。Tasgin 等人设计了适合字符串编码的单路交叉操作，通过利用 GA(genetic algorithm)算法优化社区模块度 Q 函数来实现网络最优划分的近似[16]。邓琨等人给出一种基于遗传框架的社团检测算法[17]，根据节点评价值实施有指向性的变异策略以克服随机变异的盲目性。金弟等人分析了模块性函数的局部梯度特性，结合遗传算法，提出快速有效地局部搜索变异策略，可用于大规模网络社区检测[18]。Gach 和 Hao 提出的社区检测算法[19]将遗传方法中的交叉算子和 Memetic 算法结合在一起，用 BGLL 算法产生初始解，基于优先级由两个父聚类的社区交叉产生新的社区结果。金弟等人从仿生学角度出发，以蚁群算法为框架，基于随机游走，给出一种社区发现算法[20]，将蚂蚁的局部解集成为全局解，通过“强化簇内连接，弱化簇外连接”使社区结构呈现出 来。

近两年涌现出一些比较新颖的进化社区检测算法。如 Chopade 提出了一种基于博弈论的复杂网络社区发现方法[21]，基于纳什均衡将网络划分为紧密的社区。通过重新定义的针对权重网络的节点相似性、拉普拉斯矩阵和模块度，在进化博弈中寻找针对适应度的纳什均衡点。网络中每个节点作为博弈者按照最大化收益决定将自己划分到哪个社区，直到每个节点收益不再增加，从而得到社区划分结果；段震提出基于商空间的多层粒化社区发现方法[22]，该方法对复杂网络进行多层次粒化操作，形成逐层粒化和抽象的多粒度商空间，并选择最佳粒层作为划分结果。金志刚提出基于密度峰值聚类的自适应社区发现算法（KDED）[23]，算法将节点关系量化为基于信任度的距离矩阵，根据距离矩阵核密度估计和节点影响力大小统计分析，结合热扩散模型改进计算流程，使其自适应不同规模的数据集以提高计算精度。根据密度峰值聚类原理和社区属性确定社区中心节点，然后根据节点间的距离得到社区内部层次结构和社区外部的自然结构，最后将剩余节点按距离分配到相应的社区中，得到社区划分结果。

与粒子群、蚁群等算法比较，布谷鸟算法是一种新颖的智能进化算法，具有很强的搜索能力，算法只需要一个参数，易于实现，效率高。目前，应用于模块度优化的布谷鸟算法很少见到，本文提出结合基因遗传和贪婪搜索的布谷鸟社区检测算法(cuckoo search algorithm combining gene inheritance and greedy search for community detection, GGCSCA) ，基于布谷鸟算法的框架和搜索能力，构造一种基于邻接表随机游走的全局搜索算法，并融合了全局基因保留和局部贪心算法，目的是基于智能进化获得更好的复杂网络社区识别和检测能力。实验结果说明算法有效可用，并且相对具有较好的收敛性、准确性。

2 布谷鸟算法

布谷鸟算法是 Yang Xin-she 提出的一种群智能搜索算法[24]，该算法搜索速度快，精度高，已被广泛应用于科学研究和工程实践的优化问题。该算法灵感来自于布谷鸟的寄生巢行为。每个鸟巢代表一个候选解，通过基于 $Levy$ 飞行的随机游走方式搜索新的解，具有很强的全局搜索能力。

设 $X_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$ 是某 n 维鸟巢当前解，通过式(1)产生新的解。

$$X_i^{t+1}=X_i^t+\alpha\oplus Levy \tag{1}$$

α 是步长，$Levy$ 飞行随机游走公式为

$$L(s)\sim s^{-\lambda}\left(1<\lambda\le 3\right) \tag{2}$$

其中：s 是游走步长，λ 是步长规模参数,按照 Mantegna 的算法，

$$s=\frac{\mu}{|v|^{1/\beta}} \tag{3}$$

$\mu\sim N\left(0,\sigma_\mu^2\right),v\sim N\left(0,\sigma_v^2\right)$ ◁ 是正态分布，$\sigma_v=1$，

$$\sigma_u=\left\{\frac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\Gamma\left[(1+\beta)/2\right]\beta 2^{(\beta-1)/2}}\right\}^{1/\beta} \tag{4}$$

布谷鸟算法在当前解的基础上通过 $Levy$ 飞行随机游走产生新的解，评价并保留较好的解；然后以概率 pa 丢弃部分解，用偏好搜索重新生成与丢弃解相同数量的解，再次评价并保留较好的解。

以上是基本的布谷鸟算法，适用于连续空间的函数优化问题，难以直接用于复杂网络空间的社区检测问题，本文借助于该算法的基本思想和基本框架，针对求解的问题，设计和实现了一种适合复杂网络离散空间的布谷鸟社区检测算法。

3 布谷鸟社区检测算法

3.1 评价函数

评价函数使用 Newman 的模块度函数[8,10]：

$$Q=\frac{1}{2m}\sum_{ij}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(C_i,C_j) \tag{5}$$

m 是网络的总边数，A_{ij} 是网络的邻接矩阵，k_i 是节点 i 的度，C_i 表示 i 所属的社区，当 i，j 属于同一社区时 $\delta(C_i,C_j)=1$，否则等于 0。公式的含义是：社区内总边数和网络总边数的比值减去一个期望值，该期望值假设网络是随机网络时同样的社区分配所形成的社区内总边数和网络总边数的比值大小。这是一种数学方式刻化的社区划分评价，它的取值范围是-0.5 到 1。很多社区发现算法基于模块度优化，目标是尽可能最大化模块度，但是优化 Q 是 NP 难问题，所以都是近似优化。模块度方法缺点是难以发现小于一定规模 δ 的小社团，δ 与具体的网络规模相关。尽管有这样的模块度限制[25]，模块度优化方法依然不失为一种有效地通用方法。式(5)的模块度函数针对无向无权网络，有向网络和带权网络有对应的类似模块度函数，从模块度函数优化的角度出发，本文基于无向无权网络，不失一般性。

3.2　信息编码

布谷鸟巢穴编码采用基于节点标号的编码方式，设网络为 $G=(V,E)$，节点数为 n，对于一个巢穴解 $X_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$，若 $x_{im}=k$，$m=1\cdots n$，则表示节点 m 和 k 处于同一个社区中，最后通过归并可以得到社区的划分结果。本文以邻居有序表[14]为基础，巢穴逐维在网络上进行随机游走，即 x_{im} 的值在节点 m 的邻居节点序列上随机游走取得，保障不会产生非法解。初始化时，x_{im} 随机取得节点 m 邻居节点序列中的某个节点标号。图 1(a)是简单网络示例，图 1(b)是该网络划分的两个社区，该网络邻居有序表如表 1 所示，图 1(c)中的向量表示了布谷鸟巢穴编码。

表 1　邻居有序表

节点编号	邻居节点序列			
1	2	3	--	--
2	1	3	--	--
3	1	2	4	--
4	3	5	6	--
5	4	6	7	--
6	4	5	7	--
7	5	6		

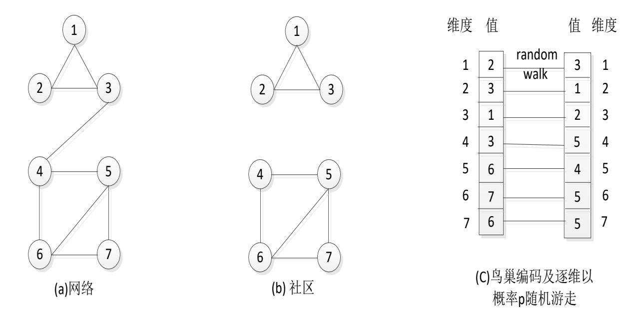


图 1　网络社区、鸟巢编码及网络上的随机游走

3.3　算法框架和实现

3.3.1 基于邻居表的逐维随机游走和基因遗传

布谷鸟搜索算法全局搜索基于随机游走。本文设计的随机游走是基于邻接有序表的逐维随机游走。某鸟巢解 t1 时刻为 $X_i^{t_1}=\{x_{i_1}^{t_1},x_{i_2}^{t_1},\cdots,x_{i_n}^{t_1}\}$，下一时刻变成 $X_i^{t_2}=\{x_{i_1}^{t_2},x_{i_2}^{t_2},\cdots,x_{i_n}^{t_2}\}$，从 $x_{ij}^{t_1}$ 到 $x_{ij}^{t_2}$ 是以概率 p 在邻接表上随机游走而得到，如图 1(c)所示。

对于节点 j，概率 $p=\frac{1}{d_j}$，d_j 是节点 j 的度。

对于每个巢穴 X_i，除了用随机游走保证解的灵活性和全局性，每一代较大程度上还要在上一代的基础上进化，以保证求解的效率和收敛。为此每一代除全局最优解外，再取前 s 个较优鸟巢解保存在精英库 $H=\{h_1,h_2,\cdots,h_s\}$。X_i 的每一位 $x_{ij}\left(j=1\cdots n\right)$ 可看做是一个基因，引入基因遗传策略：分别按概率 p_c 和 p_g 保留某精英基因(h_m)和全局最优解($gbest$)基因。每次迭代得到 x_i 的公式为：

$$x_{ij}=\begin{cases}k & rand<p_c\\ h_{mj} & p_c\le rand<p_g\\ gbest_j & rand\ge p_g\end{cases} \tag{6}$$

$$h_m=H\left(ceil(random\times s)\right) \tag{7}$$

k 由随机游走得到。式(7)表示随机抽取精英库中的某个解，$ceil$ 表示向最近大整数取整。

下面给出布谷鸟社区搜索算法框架：

算法 1　布谷鸟社区搜索算法

Input:复杂网络 $G=(V,E)$，鸟巢规模 l，偏好搜索概率 pa，迭代次数 $iter$。

Output:社区划分　$C=\{C_1,C_2,\cdots,C_k\}$

begin:

根据网络构建邻居有序表 $AdTable$；构建规模为 s 的精英库 $H=\{h_1,h_2,\cdots,h_s\}$，元素初始值为空；$t=0$；

定义目标函数为模块度函数 Q；

初始化布谷鸟巢穴种群，每个巢穴 X_i 各维随机取得邻居有序表中的值；计算目标函数值，得到初始 Q 值；

更新精英库 $H=\{h_1,h_2,\cdots,h_s\}$；

按式(6)更新巢穴，每个巢穴 X_i 部分继承优秀基因，部分在邻接表上随机游走一次，得到新的巢穴 X_i'；

根据 X_i'，计算目标函数值，若 $Q(X_i')>Q(X_i)$，则 $X_i\leftarrow X_i'$；

将 $Q(X_i)$，$i=1,2,\cdots l$ 中的最大值 $best$ 与当前整体最优 $gbest$ 比较，若 $best>gbest$，则 $gbest\leftarrow best$；

执行局部贪婪搜索算法 $GreedyLocal(X,pa)$（见算法 2），得到新的巢穴 X_i'，依次执行第 6 步，第 7 步；然后执行第 9 步；

$t=t+1$,若迭代次数 $t<iter$,转第 4 步执行;否则执行第 10 步;

退出迭代，对最优解 $gbest$ 解码，获得社区划分 $C=\{C_1,C_2,\cdots,C_k\}$。

3.3.2 局部偏好搜索

局部偏好搜索的目的是提高种群质量，加速搜索过程。本局部搜索方法使用贪婪算法[10]，利用模块度公式计算节点 i 离开当前所在社区进入其他邻接社区时产生的模块度增量 ΔQ，找到使得 ΔQ 最大的那个社区作为该节点的移入社区。这主要分两步，1）使节点 i 离开原社区成为独立节点产生的 ΔQ_1，见式(8)；2）使该独立节点 i 移入新社区产生的增量 ΔQ_2，见式(9)。$\Delta Q=\Delta Q_1+\Delta Q_2$。在式(8)和(9) 中 $\sum_{in}$ 代表节点 i 待加入（离开）的社区内部连接权重和，$\sum_{tot}$ 代表与节点 i 待加入（离开）的社区内各点连边（包括社区内连接和社区外连接）的权重之和。k_i 代表节点 i 的度，$k_{i,in}$ 代表在待加入（离开）社区内节点 i 与该社区内节点连边权重和。

$$\Delta Q_1=\left[\frac{\sum_{in}-k_{i,in}}{2m}-\left(\frac{\sum_{tot}-k_i}{2m}\right)^2-\left(\frac{k_i}{2m}\right)^2\right]$$
$$-\left[\frac{\sum_{in}}{2m}-\left(\frac{\sum_{tot}}{2m}\right)^2\right] \tag{8}$$

$$\Delta Q_2=\left[\frac{\sum_{in}+k_{i,in}}{2m}-\left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right]-$$
$$\left[\frac{\sum_{in}}{2m}-\left(\frac{\sum_{tot}}{2m}\right)^2-\left(\frac{k_i}{2m}\right)^2\right] \tag{9}$$

若 $K_i=\{k_1,k_2,\cdots,k_d\}$ 是节点 i 的邻居，则对节点 i

$$x_i'=\arg\max_j \Delta Q\left(x_i, j \mid j\in K\right) \tag{10}$$

即取使得 ΔQ 最大的邻接点作为更新位置。下面给出局部偏好搜索步骤：

算法 2 局部贪婪搜索算法

Input: 布谷鸟巢穴 $X=\{X_1,X_2,\cdots,X_l\}$，邻居有序表 $AdTable$，概率 pa

Output: 偏好搜索优化后的布谷鸟巢穴 $X'=\{X_1',X_2',\cdots,X_l'\}$

$decode(X)$，对各候选巢穴 $X_i(i=1\cdots l)$ 进行解码归并，得到对应社区划分。

以概率 pa 确定巢穴 $X_i(i=1\cdots l)$ 的部分候选基因 $x_{ij}(j=1\cdots n)$，计算当节点 j 离开所在社区加入其他邻居社区时的 ΔQ，$\Delta Q=\Delta Q_1+\Delta Q_2$ 计算见式(8)、(9)。

找到使 ΔQ 最大的社区，使 j 加入该社区。

获得新的巢穴 $X'=\{X_1',X_2',\cdots,X_l'\}$

3.4 算法时间复杂度分析

设节点数为 n，边数为 m，鸟巢数为 l，循环迭代次数为 $iter$，平均度数为 $d=\frac{m}{n}$。

算法最主要的步骤及复杂度是：a)构建邻居有序表时间复杂度为 $O(m)$；b)初始化鸟巢，复杂度为 $O(l\times n)$；c)逐维随机游走选择一个邻居节点或基因保留运算，复杂度为 $O(iter\times l\times n)$；d)解码归并社区复杂度为 $O(iter\times l\times n)$；e)设社区数是 c，计算模块度复杂性为：$O\left(iter\times l\times c\times\left(\frac{n}{c}\right)^2\right)$；f)局部贪婪搜索中计算 ΔQ 只需要局部信息，所以其复杂度为 $O(iter\times l\times d)$。复杂性最高的是 $O\left(iter\times l\times c\times\left(\frac{n}{c}\right)^2\right)$，而 $iter$，l 为常数，所以复杂度为 $O\left(\frac{n^2}{c}\right)=O(ns)$，其中 $s=\frac{n}{c}$ 是社区规模，一般远小于 n。

4 实验及分析

数值实验在 Intel i5 处理器、4G 内存和 win7 操作系统的电脑上运行，MATLAB 2011 环境下编程计算。参数设置：布谷鸟算法中，按一般性设置，种群规模 l 为 25，偏好搜索概率 pa 为 0.25；精英规模 s 为 4。

参数分析

p_c 和 p_g 分别是继承某精英基因(h_m)和全局最优解($gbest$)基因的概率阈值。以 p_g 为例，假设种群只继承最优基因，通过独立运行 20 次的平均值，得出 p_g 对模块度值的影响如图 2 所示。图中四个网络分别是 Karate（空手道俱乐部网络），Dolphin(海豚网络)，Football(足球联盟网络)[18]和 500 个节点的人工生成网络 Lfr500。可以看出，在不考虑继承其他精英基因的情况下，p_g 取 0.1 时效果最好，说明在保持一定灵活性和随机性的基础上尽量继承最优($gbest$)基因效果最佳，所以按式(6)，p_g 和 p_c 应当在 0.1-0.3 之间，且 p_c 不应该超过 p_g。综合以上分析，在本实验中 p_c 设为 0.1，p_g 取 0.2。

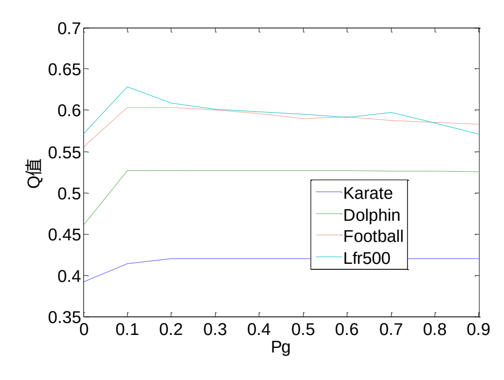


图 2 参数 p_g 对模块度值的影响

4.1 人工网络

通过 Lancichinetti 提出的基准测试网络[26]考察算法对社区的识别能力，该网络有 128 个节点，分成 4 个社区，每个社区 32 个点，节点平均度 16。生成网络时的混合参数 μ，显著影响社区结构，其作用是：以概率 μ 连接社区外部节点，以概率 1-μ 连接社区内部节点。μ 从 0.1 到 0.5 社区结构从清晰变

模糊，μ<0.5 时一个节点的邻居属于同一个社区的概率大于社区外，社区结构应该能够识别出来；当取 0.5 时，每个节点平均有一半的连接指向了社区外的节点，此时社区结构就比较模糊。

本实验采用规范化互信息(NMI)[27]来评价社区识别效果，如式(10)所示。

$$NMI(A,B)=\frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}C_{ij}\log\left(\frac{nC_{ij}}{C_iC_j}\right)}{\sum_{i=1}^{C_A}C_i\log\left(\frac{C_i}{n}\right)+\sum_{j=1}^{C_B}C_j\log\left(\frac{C_j}{n}\right)} \tag{10}$$

A 与 B 是两种社区划分，C 是混合矩阵，C_A 与 C_B 分别是 A 和 B 中社区的个数。C_{ij} 是划分 A 中的社区 i 与划分 B 中的社区 j 中重合的节点个数，C_i 表示 C 中 i 行元素之和，C_j 表示 C 中 j 列元素之和。NMI 取值范围在[0,1]区间，1 表示两社区划分完全一致，0 表示完全不同。

将算法生成的社区结构与真实社区结构进行规范化互信息评价，并与 igraph 软件中内置的 5 种社区发现经典算法比较：InfoMap，FastGreedy，LeadingEigenvector，BGLL，LPA。

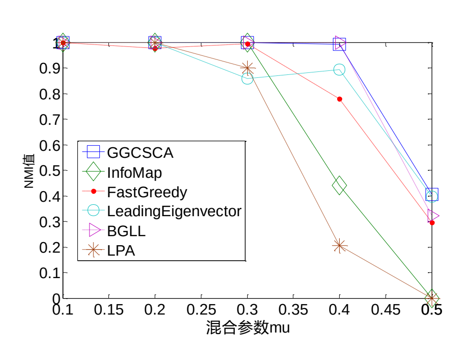


图 3　几种算法在人工网络上的 NMI 值比较

分别以混合参数 0.1-0.5 产生 5 个网络，每个算法独立运行 50 次，求得 NMI 平均值。根据图 3 可以看出，对 μ 为 0.1 和 0.2 两个网络，各个算法都能正确或基本正确地识别，μ 为 0.3 时结果产生分化，LPA 和 LeadingEigenvector 的结果分别降到了 0.9 和 0.86，到第 4 个网络只有 BGLL 和本算法 GGCSCA 能够完全正确识别，其他几个算法识别最好的 LeadingEigenvector 其 NMI 值也仅是 0.89。对前四个网络 BGLL 算法和本文算法 GGCSCA 都能准确识别，明显优于其他几种算法；对于第 5 种网络，本文算法 NMI 为 0.41 优于 BGLL 算法的 0.32。这充分证明了本算法的社区探测识别能力。

4.2　经典网络

用 3 个经典数据集[18]进行实验，分别是 Karate (空手道俱乐部网络)，Dolphin(海豚网络)，　Football(足球联盟网络)。

4.2.1 收敛情况

由图 4 可以看出算法很快就收敛。空手道俱乐部仅用 2 次迭代（一次迭代指种群的一次全局搜索和一次局部偏好搜索）就收敛到全局最优值 0.4198，迭代次数最多的 Football 迭代次数也没有超过 25 次。

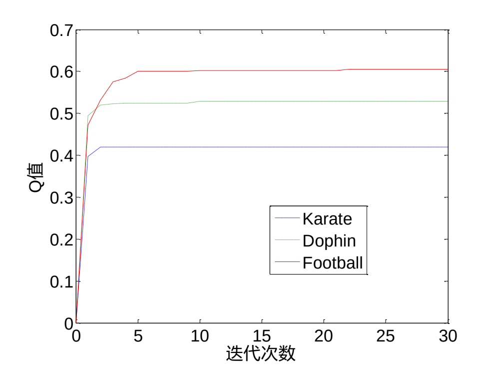


图 4　GGCSCA 算法收敛图

4.2.2 社区划分及评价

表 2 是各种算法对 3 个网络分别计算模块度 50 次得到的平均值。算法中包含两种典型智能进化算法：粒子群算法(PSO)和遗传算法 CCGA[18,28]。

表 2　各算法在经典网络上的模块度值比较

算法	Karate	Dophin	Football
BGLL	0.4176	0.5222	0.6037
CNM	0.3807	0.4955	0.5497
PSO	0.409	0.511	0.598
GN	0.4013	0.4706	0.5996
LPA	0.3805	0.4963	0.5915
CCGA	0.4198	0.5273	0.6005
infomap	0.4020	0.5236	0.6005
KDED	0.402	0.447	0.585
GGCSCA	0.4198	0.5275	0.6037

对 karate 网络，本算法 GGCSCA 模块度 Q 取得了 0.4198 的值，优于其他算法（仅有 CCGA 也取得 0.4198），高于 Karate 真实社区网络社区结构对应的 Q 值 0.3715，从模块度优化的角度来说本方法取得了很好的结果。从另外一点来看，对有些网络，数学方法衡量的模块度相对于人为划定的真实网络社区结构是有偏差的，karate 网络真实的社区结构为 2 个社区，对应于局部 Q 最大值而非全局最大值[3]，但按文献[18]说法这也是合理的，本算法将网络划分为 4 个社区，而这 4 个社区真好是 2 个真实社区的更紧凑细分。GGCSCA 算法产生的社区划分情况如图 5 与表 3 所示。

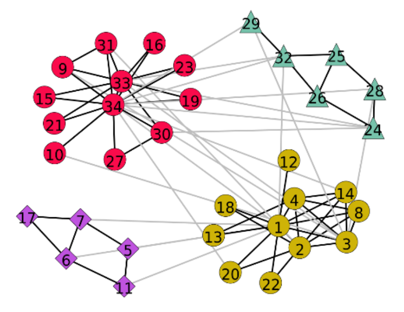


图 5　karate 网络社区图

表 3　karate 网络社区划分

真实社区		算法产生社区　（细分）	
社区编号	节点编号	社区编号	节点编号
0	1 2 3 4 5 6 7 8 11 12 13 14 17 18 20 22	0	1 2 3 4 8 12 13 14 18 20 22
		1	5 6 7 11 17
1	9 10 15 16 19 21 23 24 25 26 27 28 29 30 31 32 33 34	2	9 10 15 16 19 21 23 27 30 31 33 34
		3	24 25 26 28 29 32

与空手道网络类似，海豚网络的真实网络对应于 Q 值 0.3722，也是收敛于局部最优。本算法得到平均值 0.5278，优于其他几个算法，模块度优化取得了好的结果。真实网络划分为 2 个社区，本方法所得是 5 个更紧凑划分的社区，一个社区完全对应，另一个社区细分为 4 个社区，划分社区情况如图 6 与表 4 所示。

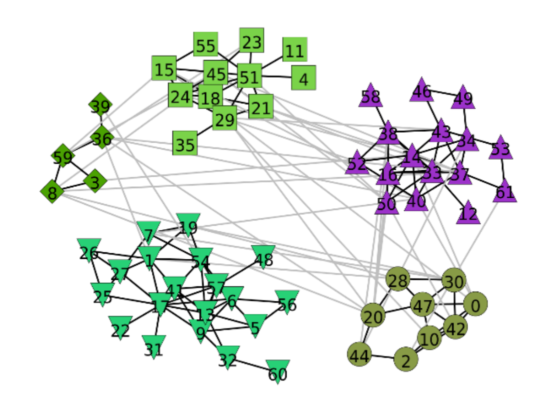


图 6　Dolphin 网络社区图

表 4　Dolphin 网络社区划分

真实社区		算法产生社区　（细分）	
社区编号	节点编号	社区编号	节点编号
0	1 5 6 7 9 13 17 19 22 25 26 27 31 32 41 48 54 56 57 60	1	1 5 6 7 9 13 17 19 22 25 26 27 31 32 41 48 54 56 57 60
1	0 2 3 4 8 10 11 12 14 15 16 18 20 21 23 24 28 29 30 33 34 35 36 37 38 39 40 42 43 44 45 46 47 49 50 51 52 53 55 58 59 61	4	0 2 10 20 28 30 42 44 47
		2	4 11 15 18 21 23 24 29 35 45 51 55
		0	3 8 36 39 59
		3	12 14 16 33 34 37 38 40 43 46 49 50 52 53 58 61

对于 Football 网络，本算法 GGCSCA 计算的 Q 平均值取得 0.6037，与 BGLL 算法的值相同，优于其他算法。足球俱乐部的真实网络划分为 12 个社区，本算法运行过程中最好的划分结果为 11 个社区，有 6 个社区完全正常匹配， 5 个社区的节点基本被正确识别，1 个真实社区节点被分配到其他社区，如图 7 和表 5 所示（上标代表该节点在对方社区的编号）。

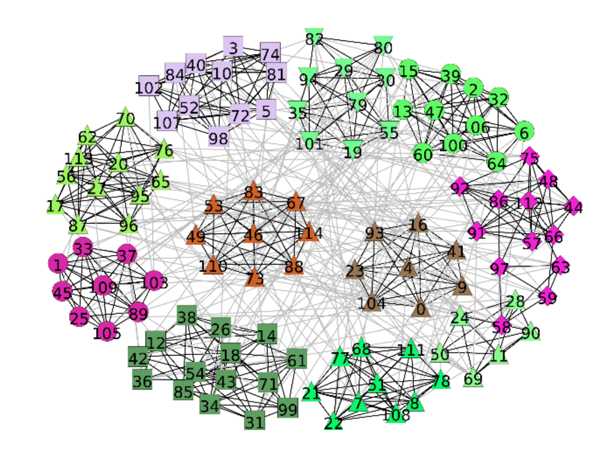


图 7　Football 网络社区图

表 5　Football 网络社区划分

真实社区		算法产生社区　（匹配）	
社区编号	节点编号	社区编号	节点编号
0	1 25 33 37 45 89 103 105 109	4	1 25 33 37 45 89 103 105 109
1	19 29 30 35 55 79 94 101	3	19 29 30 35 55 79 $(80\ 82)^{11}$ 94 101
2	2 6 13 15 32 39 47 60 64 100 106	0	2 6 13 15 32 39 47 60 64 100 106
3	3 5 10 40 52 72 74 81 84 98 102 107	9	3 5 10 40 52 72 74 81 84 98 102 107
4	44 48 57 66 75 86 91 92 110^{6} 112	8	44 48 57 58^{10} $(59\ 63)^{9}$66 75 86 91 92 97^{9} 112
5	12 14 18 26 31 34 38 43 54 61 71 85 99	1	12 14 18 26 31 34 36^{11} 38 42^{11} 43 54 61 71 85 99
6	0 4 9 16 23 41 93 104	2	0 4 9 16 23 41 93 104
7	7 8 21 22 51 68 77 78 108 111	7	7 8 21 22 51 68 77 78 108 111
8	17 20 27 56 62 65 70 76 87 95 96 113	5	17 20 27 56 62 65 70 76 87 95 96　113
9	11 24 50 $(59\ 63)^{8}$ 69 97^{8}	10	11 24 28^{10} 50 69 90^{11}
10	28^{10} 46 49 53 58^{8} 67 73 83 88 114	6	49 53 67 73 83 88 110^{4} 114
11	$(36\ 42)^{1}$ $(80\ 82)^{3}$ 90^{10}		

5　结束语

本文提出的算法以模块度 Q 为评价函数，以布谷鸟算法为框架，结合邻接表上的随机游走和基因遗传策略，并应用最大模块度增量的局部偏好搜索，在保证全局灵活性的同时加快种

群的收敛。在基准网络和真实网络上的实验说明本算法具有较好的社区识别和检测能力。算法无须提供有关社区的先验知识和假设，可以有效揭示网络内在的社区结构。下一步研究将其与其他算法结合，提高搜索效率；用并行化机制实现，应用到银行交易网络等大型实际复杂网络的社区发掘研究上。

参考文献：

[1] Girvan M, Newman M E J. Community Structure in Social And Biological Networks [J]. Proceedings of National Academy of Sciences of the United States of America, 2002, 99 (12): 7821-7826.

[2] Fortunato S. Community Detection in Graphs [J]. Physics Reports, 2010, 486 (3-5): 75-174.

[3] 杨博，刘大有，金弟，等. 复杂网络聚类方法 [J]. 软件学报, 2009, 20 (1): 54-58.

[4] White S, Smyth P. A Spectral Clustering Approach to Finding Communities in Graphs [C]// Proc of SIAM International Conference on Data Mining. 2005: 76−84.

[5] Newman M E J. Fast Algorithm for Detecting Community Structure in Networks [J]. Physical Review E. 2004, 69 (6): 066133.

[6] Clauset A, Newman M E J, Moore C. Finding Community Structure in Very Large Networks [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics [J]. 2004, 70 (2): 066111.

[7] Pons P, Latapy M. Computing Communities in Large Networks Using Random Walks [J]. Journal of Graph Algorithms and Applications . 2006, 10 (2): 191-218.

[8] Newman M E J, Girvan M. Finding and Evaluating Community Structure in Networks [J]. Physical Review E. 2004, 69 (2): 026113.

[9] Guimera R, Amaral L A N. Functional Cartography of Complex Metabolic Networks [J]. Natrue, 2005, 433 (7028): 895-900.

[10] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast Unfolding of Community Hierarchies in Large Networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10: 10008.

[11] Duch J, Arenas A. Community Detection in Complex Networks Using Extreme Optimization [J]. Physical Review E, 2005, 72 (2): 027104.

[12] 黄发良，张师超，朱晓峰. 基于多目标优化的网络社区发现方法 [J]. 软件学报, 2013, 24 (9): 2062-2077.

[13] Zhou X, Liu Y H, Li B. A Multi-Objective Discrete Cuckoo Search Algorithm with Local Search for Community Detection In Complex Networks [J] , Modern Physics Letters B, 2016, 30 (7): 1-20.

[14] 黄发良，肖南峰. 网络社区发现的粒子群优化算法 [J]. 控制理论与应用, 2011, 28 (9): 1135-1139.

[15] 邱晓辉，陈羽中. 一种面向社会网络社区发现的改进粒子群优化算法 [J]. 小型微型计算机系统, 2014, 35 (6): 1422-1425.

[16] Tasgin M, Herdagdelen A, Bingol H. Community Detection in Complex Networks Using Genetic Algorithms [EB/OL]. (2007-11-04) [2017-05-20]. https://arxiv. org/pdf/0711. 0491. pdf.

[17] 邓琨，张健沛，杨静. 利用改进遗传算法进行复杂网络社团发现 [J]. 哈尔滨工程大学学报, 2013, 34 (11): 1438-1443.

[18] 金弟，刘杰，杨博，等. 局部搜索与遗传算法结合的大规模复杂网络社区探测 [J]. 自动化学报, 2011, 37 (7): 873-879.

[19] Gach O, Hao J K. A Memetic Algorithm for Community Detection in Complex Networks [C]// Proc of International Conference on Parallel Problem Solving From Nature. [S. l.] : Springer-Verlag, 2012: 327-336.

[20] 金弟，杨博，刘杰，等. 复杂网络簇结构探测——基于随机游走的蚁群算法 [J]. 软件学报, 2012, 23 (3): 451-456.

[21] Chopade P. A Framework for community detection in large networks using game-theoretic modeling [J]. Ieee Transactions On Big Data, 2017, 3 (3): 276-287.

[22] 段震，闵星，王倩倩，等. 基于商空间的多层粒化社区发现方法 [J]. 南京大学学报：自然科学版, 2017, 53 (4): 764-772.

[23] 金志刚，徐珮轩. 密度峰值聚类的自适应社区发现算法 [J/OL]. 哈尔滨工业大学学报，2018，50 (5) . [2017-08-24]. http://kns. cnki. net/kcms/detail/23. 1235. T. 20170824. 1116. 002. html.

[24] Yang X S. Nature Inspired Meta heuristic Algorithms [M]. 2nd ed. [S. l.] : Luniver Press, 2010.

[25] Fortunato S, Barthelemy M. Resolution Limit In Community Detection [J]. Proceedings of National Academy of Sciences of the United States of America, 2007, 104 (1): 36-41.

[26] Lancichinetti A, Fortunato S, Radicchi F. Benchmark Graphs for Testing Community Detection Algorithms [J]. Physical Review E, 2008, 78 (4): 046110.

[27] Danon L, Duch J, Diaz-Guilera A, et al. Comparing Community Structure Identication [J]. Journal of Statistical Mechanics: Theory and Experiment, 2005, 78 (9): 09008.

[28] 何东晓，周栩，王佐，等. 复杂网络社区挖掘——基于聚类融合的遗传算法 [J]. 自动化学报, 2010, 36 (8): 1160-1170.